Electronics: active circuit elements

Module by: Free High School Science Texts Project.

Active Circuit Elements

The components you have been learning about so far — resistors, capacitors and inductors — are called **passive** components. They do not change their behaviour or physics and therefore always have the same response to changes in voltage or current. **Active** components are quite different. Their response to changes in input allows them to make amplifiers, calculators and computers.

The Diode

A diode is an electronic device that allows current to flow in one direction only.

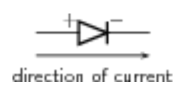

Figure 1: Diode circuit symbol and direction of flow of current.

A diode consists of two doped semi-conductors joined together so that the resistance is low when connected one way and very high the other way.

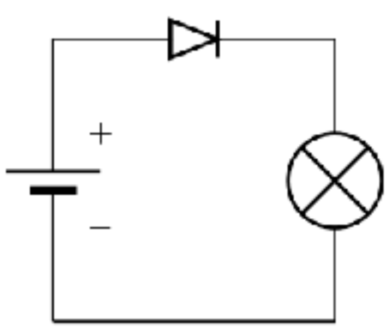

Figure 2: Operation of a diode. (Left) The diode is forward biased and current is permitted. The negative terminal of the battery is connected to the negative terminal of the diode. (Right) The diode is reverse biased and current flow is not allowed. The negative terminal of the battery is connected to the positive terminal of the diode.

A full explanation of diode operation is complex. Here is a simplified description. The diode consists of two semiconductor blocks attached together. Neither block is made of pure silicon — they are both **doped**. Doping was described in more detail in Section 10.3.

In short, p-type semiconductor has fewer free electrons than normal semiconductor. `P' stands for `positive', meaning a lack of electrons, although the material is actually neutral. The locations where electrons are missing are called **holes**. This material can conduct electricity well, because electrons can move into the holes, making a new hole somewhere else in the material. Any extra electrons introduced into this region by the circuit will fill some or all of the holes.

In n-type semiconductor, the situation is reversed. The material has more free electrons than normal semiconductor. `N' stands for `negative', meaning an excess of electrons, although the material is actually neutral.

When a p-type semiconductor is attached to an n-type semiconductor, some of the free electrons in the n-type move across to the p-type semiconductor. They fill the available holes near the junction. This means that the region of the n-type semiconductor nearest the junction has no free electrons (they've moved across to fill the holes). This makes this n-type semiconductor positively charged. It used to be electrically neutral, but has since lost electrons.

The region of p-type semiconductor nearest the junction now has no holes (they've been filled in by the migrating electrons). This makes the p-type semiconductor negatively charged. It used to be electrically neutral, but has since gained electrons.

Without free electrons or holes, the central region can not conduct electricity well. It has a high resistance, and is called the **depletion band**. This is shown in **Figure 3** (#uid32) .

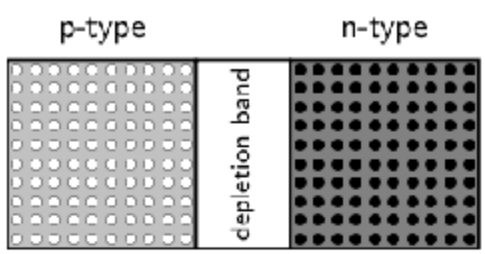

Figure 3: A diode consists of two doped semi-conductors joined together so that the resistance is low when connected one way and very high the other way.

You can explain the high resistance in a different way. A free electron in the n-type semiconductor will be repelled from the p-type semiconductor because of its negative charge. The electron will not go into the depletion band, and certainly won't cross the band to the p-type semiconductor. You may ask, “But won't a free electron in the p-type semiconductor be attracted across the band, carrying a current?” But there are no free electrons in p-type semiconductor, so no current of this kind can flow.

If the diode is reverse-biased, the $+$ terminal of the battery is connected to the n-type semiconductor. This makes it even more negatively charged. It also removes even more of the free electrons near the depletion band. At the same time, the $-$ terminal of the battery is connected to the p-type silicon. This will supply free electrons and fill in more of the holes next to the depletion band. Both processes cause the depletion band to get wider. The resistance of the diode (which was already high) increases. This is why a reverse-biased diode does not conduct.

Another explanation for the increased resistance is that the battery has made the p-type semiconductor

more negative than it used to be, making it repel any electrons from the n-type semiconductor which attempt to cross the depletion band.

On the other hand, if the diode is forward biased, the depletion band is made narrower. The negative charge on the p-type silicon is cancelled out by the battery. The greater the voltage used, the narrower the depletion band becomes. Eventually, when the voltage is about 0,6 V (for silicon) the depletion band disappears. Once this has occurred, the diode conducts very well.

The Diode

1. What is a diode?
2. What is a diode made of?
3. What is the term which means that a diode is connected the `wrong way' and little current is flowing?
4. Why is a diode able to conduct electricity in one direction much more easily than the other?

The Light-Emitting Diode (LED)

A light-emitting diode (LED) is a diode device that emits light when charge flows in the correct direction through it. If you apply a voltage to force current to flow in the direction the LED allows, it will light up.

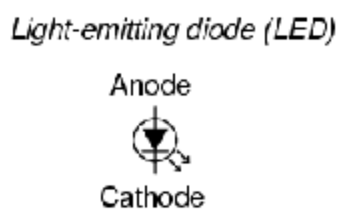

Figure 4: Symbol for a light-emitting diode with anode and cathode labeled.

Circuit Symbols

This notation of having two small arrows pointing away from the device is common to the schematic symbols of all light-emitting semiconductor devices. Conversely, if a device is light-activated (meaning that incoming light stimulates it), then the symbol will have two small arrows pointing toward it. It is interesting to note, though, that LEDs are capable of acting as light-sensing devices: they will generate a small voltage when exposed to light, much like a solar cell on a small scale. This property can be gainfully applied in a variety of light-sensing circuits.

The colour depends on the semiconducting material used to construct the LED, and can be in the near-ultraviolet, visible or infrared part of the electromagnetic spectrum.

NOTE: Interesting Fact :

Nick Holonyak Jr. (1928) of the University of Illinois at Urbana-Champaign developed the first practical visible-spectrum LED in 1962.

Light emission

The wavelength of the light emitted, and therefore its colour, depends on the materials forming the p-n junction. A normal diode, typically made of silicon or germanium, emits invisible far-infrared light (so it can't be seen), but the materials used for an LED can emit light corresponding to near-infrared, visible or near-ultraviolet frequencies.

LED applications

LEDs have many uses. Some of these are given here.

1. thin, lightweight message displays, e.g. in public information signs (at airports and railway stations, among other places)
2. status indicators, e.g. on/off lights on professional instruments and consumers audio/video equipment
3. infrared LEDs in remote controls (for TVs, VCRs, etc.)
4. clusters of LEDs are used in traffic signals, replacing ordinary bulbs behind coloured glass
5. car indicator lights and bicycle lighting
6. calculator and measurement instrument displays (seven segment displays), although now mostly replaced by LCDs
7. red or yellow LEDs are used in indicator and [alpha]numeric displays in environments where night vision must be retained: aircraft cockpits, submarine and ship bridges, astronomy observatories, and in the field, e.g. night time animal watching and military field use
8. red or yellow LEDs are also used in photographic darkrooms, for providing lighting which does not lead to unwanted exposure of the film
9. illumination, e.g. flashlights (a.k.a. torches, UK), and backlighting for LCD screens
10. signaling/emergency beacons and strobes
11. movement sensors, e.g. in mechanical and optical computer mice and trackballs
12. in LED printers, e.g. high-end colour printers

LEDs offer benefits in terms of maintenance and safety.

1. The typical working lifetime of a device, including the bulb, is ten years, which is much longer than the lifetimes of most other light sources.
2. LEDs fail by dimming over time, rather than the abrupt burn-out of incandescent bulbs.
3. LEDs give off less heat than incandescent light bulbs and are less fragile than fluorescent lamps.
4. Since an individual device is smaller than a centimetre in length, LED-based light sources used for illumination and outdoor signals are built using clusters of tens of devices.

Because they are monochromatic, LED lights have great power advantages over white lights where a specific colour is required. Unlike the white lights, the LED does not need a filter that absorbs most of the emitted white light. Coloured fluorescent lights are made, but they are not widely available. LED lights are inherently coloured, and are available in a wide range of colours. One of

the most recently introduced colours is the emerald green (bluish green, wavelength of about 500 nm) that meets the legal requirements for traffic signals and navigation lights.

> **NOTE: Interesting Fact :**
>
> The largest LED display in the world is 36 m high, at Times Square, New York, U.S.A.

There are applications that specifically require light that does not contain any blue component. Examples are photographic darkroom safe lights, illumination in laboratories where certain photo-sensitive chemicals are used, and situations where dark adaptation (night vision) must be preserved, such as cockpit and bridge illumination, observatories, etc. Yellow LED lights are a good choice to meet these special requirements because the human eye is more sensitive to yellow light.

The Light Emitting Diode

1. What is an LED?
2. List 5 applications of LEDs.

Transistor

The diode is the simplest semiconductor device, made up of a p-type semiconductor and an n-type semiconductor in contact. It can conduct in only one direction, but it cannot control the size of an electric current. Transistors are more complicated electronic components which can control the size of the electric current flowing through them.

This enables them to be used in amplifiers. A small signal from a microphone or a radio antenna can be used to control the transistor. In response, the transistor will then increase and decrease a much larger current which flows through the speakers.

> **NOTE: Interesting Fact :**
>
> One of the earliest popular uses of transistors was in cheap and portable radios. Before that, radios were much more expensive and contained glass valves which were fragile and needed replacing. In some parts of the world you can still hear people talking about their `transistor' — they mean their portable radio.

You can also use a small current to turn the transistor on and off. The transistor then controls a more complicated or powerful current through other components. When a transistor is used in this way it is said to be in **switching** mode as it is acting as a remotely controlled switch. As we shall see in the final sections of this chapter, switching circuits can be used in a computer to process and store digital information. A computer would not work without the millions (or billions) of transistors in it.

There are two main types of transistor - bipolar transistors (NPN or PNP), and field effect transistors (FETs). Both use doped semiconductors, but in different ways. You are mainly required to know about field effect transistors (FETs), however we have to give a brief description of bipolar transistors so that you see the difference.

Bipolar Transistors

Bipolar transistors are made of a doped semiconductor `sandwich'. In an NPN transistor, a very thin layer of p-type semiconductor is in between two thicker layers of n-type semiconductor. This is shown in **Figure 5** (#uid61) . Similarly an PNP transistor consists of a very thin n-type layer in between two thicker layers of p-type semiconductor.

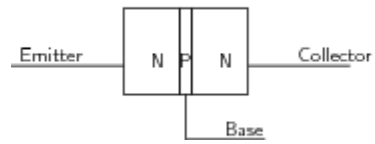

Figure 5: An NPN transistor. This is a type of bipolar transistor.

In an NPN transistor a small current of electrons flows from the emitter (E) to the base (B). Simultaneously, a much larger current of electrons flows from the emitter (E) to the collector (C). If you lower the number of electrons able to leave the transistor at the base (B), the transistor automatically reduces the number of electrons flowing from emitter (E) to collector (C). Similarly, if you increase the current of electrons flowing out of the base (B), the transistor automatically also increases the current of electrons flowing from emitter (E) to collector (C). The transistor is designed so that the current of electrons from emitter to collector (I_{EC}) is proportional to the current of electrons from emitter to base (I_{EB}). The constant of proportionality is known as the **current gain**β. So $I_{EC} = \beta I_{EB}$.

How does it do it? The answer comes from our work with diodes. Electrons arriving at the emitter (n-type semiconductor) will naturally flow through into the central p-type since the base-emitter junction is forward biased. However if none of these electrons are removed from the base, the electrons flowing into the base from the emitter will fill all of the available `holes'. Accordingly, a large depletion band will be set up. This will act as an insulator preventing current flow into the collector as well. On the other hand, if the base is connected to a positive voltage, a small number of electrons will be removed by the base connection. This will prevent the `holes' in the base becoming filled up, and no depletion band will form. While some electrons from the emitter leave via the base connection, the bulk of them flow straight on to the collector. You may wonder how the electrons get from the base into the collector (it seems to be reverse biased). The answer is complicated, but the important fact is that the p-type layer is extremely thin. As long as there is no depletion layer, the bulk of the electrons will have no difficulty passing straight from the n-type emitter into the n-type collector. A more satisfactory answer can be given to a university student once band theory has been explained.

Summing up, in an NPN transistor, a small flow of electrons from emitter (E) to base (B) allows a much larger flow of electrons from emitter (E) to collector (C). Given that conventional current (flowing from $+$ to $-$) is in the opposite direction to electron flow, we say that a small conventional current from base to emitter allows a large current to flow from collector to emitter.

A PNP transistor works the other way. A small conventional current from emitter to base allows a much larger conventional current to flow from emitter to collector. The operation is more complicated to explain since the principal charge carrier in a PNP transistor is not the electron but the `hole'.

The operation of NPN and PNP transistors (in terms of conventional currents) is summarized in **Figure 6** (#uid62) .

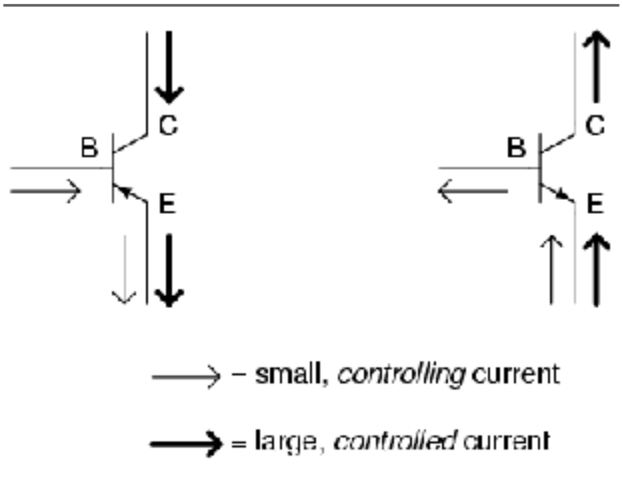

Figure 6: An overview of bipolar transistors as current amplifiers. (Left) An NPN transistor. (Right) A PNP transistor.

> **NOTE: Interesting Fact :**
>
> The transistor is considered by many to be one of the greatest discoveries or inventions in modern history, ranking with banking and the printing press. Key to the importance of the transistor in modern society is its ability to be produced in huge numbers using simple techniques, resulting in vanishingly small prices. Computer "chips" consist of millions of transistors and sell for Rands, with per-transistor costs in the thousandths-of-cents. The low cost has meant that the transistor has become an almost universal tool for non-mechanical tasks. Whereas a common device, say a refrigerator, would have used a mechanical device for control, today it is often less expensive to simply use a few million transistors and the appropriate computer program to carry out the same task through "brute force". Today transistors have replaced almost all electromechanical devices, most simple feedback systems, and appear in huge numbers in everything from computers to cars.

> **NOTE: Interesting Fact :**
>
> The transistor was invented at Bell Laboratories in December 1947 (first demonstrated on December 23) by John Bardeen, Walter Houser Brattain, and William Bradford Shockley, who were awarded the Nobel Prize in physics in 1956.

The Field Effect Transistor (FET)

To control a bipolar transistor, you control the **current** flowing into or out of its base. The other type of transistor is the field effect transistor (FET). FETs work using control **voltages** instead. Accordingly they can be controlled with much smaller currents and are much more economic to use.

NOTE: Interesting Fact :

No-one would build a computer with billions of bipolar transistors — the current in each transistor's base might be small, but when you add up all of the base currents in the millions of transistors, the computer as a whole would be consuming a great deal of electricity and making a great deal of heat. Not only is this wasteful, it would prevent manufacturers making a computer of convenient size. If the transistors were too close together, they would overheat.

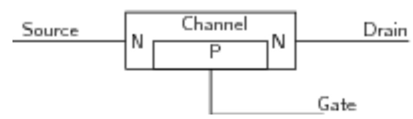

Figure 7

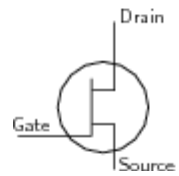

Figure 8

A field effect transistor (FET). The diagram on the top shows the semiconductor structure. The diagram underneath shows its circuit symbol.

The three terminals of the FET are called the *source* (S), *drain* (D) and *gate* (G), as shown in **(Reference)** (). When the gate is not connected, a current of electrons can flow from source (S) to drain (D) easily along the channel. The source is, accordingly, the negative terminal of the transistor. The drain, where the electrons come out, is the positive terminal of the transistor. A few electrons will flow from the n-type channel into the p-type semiconductor of the gate when the device is manufactured. However, as these electrons are not removed (the gate is not connected), a depletion band is set up which prevents further flow into the gate.

In operation, the gate is connected to negative voltages relative to the source. This makes the p-n junction between gate and channel reverse-biased. Accordingly no current flows from the source into the gate. When the voltage of the gate is lowered (made more negative), the depletion band becomes wider. This enlarged depletion band takes up some of the space of the channel. So the lower the voltage of the gate (the more negative it is relative to the source), the larger the depletion band. The larger the depletion band, the narrower the channel. The narrower the channel, the harder it is for electrons to flow from source to drain.

The voltage of the gate is not the only factor affecting the current of electrons between the source and the drain. If the external circuit has a low resistance, electrons are able to leave the drain easily. If the external circuit has a high resistance, electrons leave the drain slowly. This creates a kind of `traffic jam' which slows the passage of further electrons. In this way, the voltage of the drain regulates itself, and is more or less independent of the current demanded from the drain.

Once these two factors have been taken into account, it is fair to say that the positive output voltage (the voltage of the drain relative to the source) is proportional to the negative input voltage (the

voltage of the gate relative to the source).

For this reason, the field effect transistor is known as a voltage amplifier. This contrasts with the bipolar transistor which is a current amplifier.

Field Effect Transistors

1. What are the two types of bipolar transistor? How does their construction differ?
2. What are the three connections to a bipolar transistor called?
3. Why are very few electrons able to flow from emitter to collector in an NPN transistor if the base is not connected?
4. Why do you think a bipolar transistor would not work if the base layer were too thick?
5. "The bipolar transistor is a current amplifier." What does this statement mean?
6. Describe the structure of a FET.
7. Define what is meant by the source, drain and gate. During normal operation, what will the voltages of drain and gate be with respect to the source?
8. Describe how a depletion layer forms when the gate voltage is made more negative. What controls the width of the depletion layer?
9. "The field effect transistor is a voltage amplifier." What does this statement mean?
10. The amplifier in a cheap radio will probably contain bipolar transistors. A computer contains many field effect transistors. Bipolar transistors are more rugged and less sensitive to interference than field effect transistors, which makes them more suitable for a simple radio. Why are FETs preferred for the computer?

The Operational Amplifier

The operational amplifier is a special kind of voltage amplifier which is made from a handful of bipolar or field effect transistors. Operational amplifiers are usually called **op-amps** for short. They are used extensively in all kinds of audio equipment (amplifiers, mixers and so on) and in instrumentation. They also have many other uses in other circuits - for example comparing voltages from sensors.

Operational amplifiers are supplied on Integrated Circuits (I.C.s). The most famous operational amplifier I.C. is numbered 741 and contains a single operational amplifier on an integrated circuit (`chip') with eight terminals. Other varieties can be bought, and you can get a single integrated circuit with two or four `741'-type operational amplifiers on it.

The symbol for an op-amp is shown in **Figure 9** (#uid75). The operational amplifier has two input terminals and one output terminal. The voltage of the output terminal is proportional to the difference in voltage between the two input terminals. The output terminal is on the right (at the sharp point of the triangle). The two input terminals are drawn on the left. One input terminal (labelled with a $+$ on diagrams) is called the **non-inverting input**. The other input terminal (labelled $-$) is called the **inverting input**. The labels $+$ and $-$ have nothing to do with the way in which the operational amplifier is connected to the power supply. Operational amplifiers must be connected to the power supply, but this is taken for granted when circuit diagrams are drawn, and these connections are not shown on circuit diagrams. Usually, when drawing electronic circuits, `0V' is taken to mean the negative terminal of the power supply. This is not the

case with op-amps. For an op-amp, `0V' refers to the voltage midway between the $+$ and $-$ of the supply.

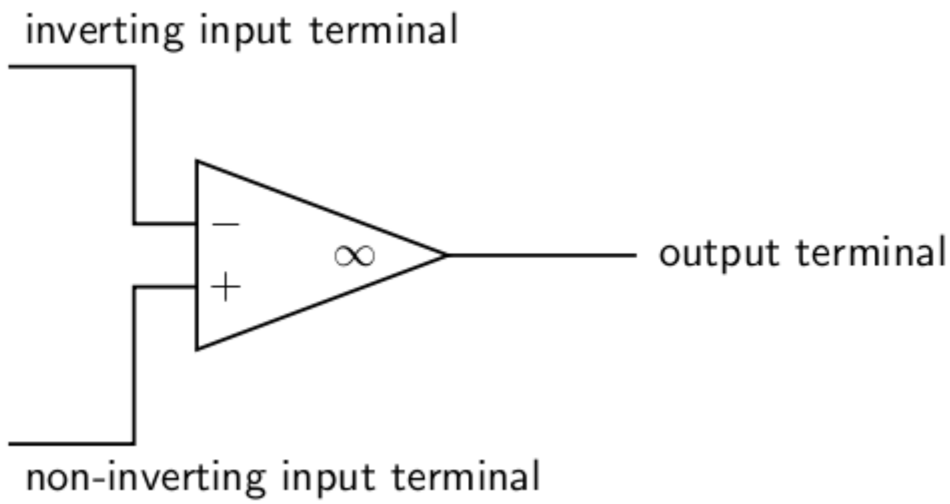

Figure 9: Circuit symbol for an operational amplifier. The amplifier must also be connectd to the $+$ and $-$ terminals of the power supply. These connections are taken for granted and not shown.

The output voltage of the amplifier V_{out} is given by the formula $V_{out} = A(V_+ - V_-)$ where A is a constant called the **open loop gain**, and V_+ and V_- are the voltages of the two input terminals. That said, the output voltage can not be less than the voltage of the negative terminal of the battery supplying it or higher than the positive terminal of the battery supplying it. You will notice that V_{out} is positive if $V_+ > V_-$ and negative if $V_+ < V_-$. This is why the $-$ input is called the inverting input: raising its voltage causes the output voltage to *drop*.

The input resistance of an operational amplifier is very high. This means that very little current flows into the input terminals during operation.

If all of the transistors in the operational amplifier were identical then the output voltage would be zero if the two inputs were at equal voltages. In practice this is not quite the case, and for sensitive work a **trimming potentiometer** is connected. This is adjusted until the op-amp is zeroed correctly.

Simple operational amplifiers require the trimming potentiometer to be built into the circuit containing them, and an example is shown in [Figure 10 (#uid76)]. Other operational amplifier designs incorporate separate terminals for the trimming potentiometer. These special terminals are labelled **offset** on the manufacturer's diagram. The exact method of connecting the potentiometer to the offset terminals can depend on the design of the operational amplifier, and you need to refer to the manufacturer's data sheet for details of which potentiometer to use and how to connect it.

For most commercially produced operational amplifiers (known as op-amps for short), the open loop gain A is very large and does not stay constant. Values of 100 000 are typical. Usually a designer would want an amplifier with a stable gain of smaller value, and builds the operational amplifier into a circuit like the one in [Figure 10 (#uid76)].

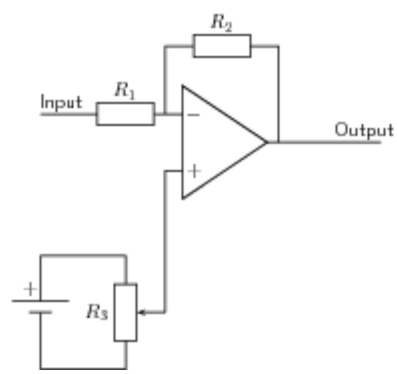

Figure 10: An inverting amplifier built using an operational amplifier. The connections from battery to operational amplifier are not shown. The output voltage $V_{out} = -R_2 V_{in}/R_1$, as explained in the text. The potentiometer R_3 is a trimming potentiometer. To set it, the input is connected to zero volts. The trimming potentiometer is then adjusted until $V_{out} = 0$. In all operational amplifier circuits, zero volts is midway between the $+$ and $-$ of the supply.

Calculating the gain of the amplifier in Figure 10 (#uid76) .

1. The input resistance of the operational amplifier is very high. This means that very little current flows into the inverting input of the op-amp. Accordingly, the current through resistor R_1 must be almost the same as the current through resistor R_2. This means that the ratio of the voltage across R_1 to the voltage across R_2 is the same as the ratio of the two resistances.
2. The open loop gain A of the op-amp is very high. Assuming that the output voltage is less than a few volts, this means that the two input terminals must be at very similar voltages. We shall assume that they are at the same voltage.
3. We want the output voltage to be zero if the input voltage is zero. Assuming that the transistors within the op-amp are very similar, the output voltage will only be zero for zero input voltage if V_+ is very close to zero. We shall assume that $V_+ = 0$ when the trimming potentiometer is correctly adjusted.
4. It follows from the last two statements that $V_- \approx 0$, and we shall assume that it is zero.
5. With these assumptions, the voltage across R_2 is the same as V_{out}, and the voltage across R_1 is the same as V_{in} . Since both resistors carry the same current (as noted in point 1), we may say that the magnitude of $V_{out}/V_{in} = R_2/R_1$. However, if V_{in} is negative, then V_{out} will be positive. Therefore it is customary to write the gain of this circuit as $V_{out}/V_{in} = -R_2/R_1$.

Operational Amplifiers

1. What are operational amplifiers used for?
2. Draw a simple diagram of an operational amplifier and label its terminals.
3. Why is a trimming potentiometer needed when using an op-amp?

More about this module: Metadata | Downloads | Version History

- How to **reuse** and attribute this content
- How to **cite** and attribute this content

Last edited by Free High School Science Texts Project on Aug 3, 2011 5:17 am GMT-5.